DOCUMENT RESUME

ED 220 111 IR 010 332

AUTHOR Batty, David; Hitchens, Christopher
TITLE Information System Design Based on User Concept Systems and User Vocabulary.
PUB DATE May 81
NOTE 12p.; Paper presented at the Mid-Year Meeting of the American Society for Information Science (10th, Durango, CO, May 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Oriented Programs; *Indexing; Information Retrieval; *Information Systems; *Lexicography; *Subject Index Terms; Systems Development; *Vocabulary; Word Frequency
IDENTIFIERS *Users

ABSTRACT

Recent work on the development of Synthesized User Based Terminology Index Languages (SUBTIL), i.e., index and query languages for information systems which employ vocabulary elicited from potential users of the system, is described in this paper. Following a brief review of the literature on thesaurus construction and index languages is a discussion of a proposed procedure for eliciting a common and significant indexing language from a user group. The role of a computer in the procedure is described. Preliminary tests of the procedure are then summarized and the functions performed by a set of computer programs for analyzing the results of vocabulary elicitation are reviewed. Some of the questions which remain to be answered regarding the utility of SUBTIL are outlined, and the potential advantages of the SUBTIL process are considered. A 12-item reference list accompanies the text. (Author/JL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220111

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

INFORMATION SYSTEM DESIGN BASED ON USER CONCEPT SYSTEMS AND USER VOCABULARY

by

David Batty

and

Christopher Hitchens

IR010332

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

David Batty

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

(In Using information. Papers contributed to the tenth mid-year meeting of the American Society for Information Science, Durango, May 1981. ASIS, 1981. Part One: p. 1; Part Two: pp. 2-11.)

INFORMATION SYSTEM DESIGN BASED ON USER CONCEPT SYSTEMS AND USER VOCABULARY

David Batty, Alpha Omega Group
and
Christopher Hitchens, MAXIMA Corporation

Contemporary thesaurus construction methods are based on the subject literature. This may not be appropriate for inter- and multi-disciplinary fields, where no coherent body of literature exists. A method is proposed to use the language of the potential user group to develop an index or query language for their needs. A sequence of individual user responses to trigger terms, analyzed at the group level for common or significantly related terms, yields high-relevance terms for use as trigger terms in further rounds. It should be possible to develop a working vocabulary for indexing or searching of several hundred terms, in a few hours of interaction for each user.

The present paper describes further work on SUBTIL (Synthesized User Based Terminology Index Languages), a system described in earlier reports by one of the authors (1,2).

The construction and maintenance of index languages have been based traditionally on a document collection. The Library of Congress used the 'literary warrant' of its own collection in the development of its classification; the content and structure of the ERIC Thesaurus depends largely on the ERIC data base. Methods of this kind generally analyze a sample of the document base for common themes, concepts, and terminology. From the sample a prototype language is developed, which is then tested (though not always rigorously) on more documents. The language then grows by the addition of more terms and references as more documents are collected that cannot be described by the existing language. If the growing language is tested by means of a Cranfield-type test, some user queries are used as a test base, but these too are expressed in terms of the document-derived vocabulary.

However, for the last thirty years or more there has been a minor, but steady and increasingly significant interest in user vocabulary and syntax as a basis for index language development. Mooers (3) showed a sample document set to engineers to establish a vocabulary meaningful to them. Bourne and others (4) in 1961 advocated interviews to establish user requirements and characteristics. Greer reported in 1962 (5) and 1965 (6) on interviews to elicit typical questions that the documents might answer and therefore user-oriented terms with which to index the literature. However, all these studies used explicit language and labels in the context of the documents. In 1966 Reiser (7) described the evaluation of a conventionally developed but growing thesaurus by interacting with its users.

But something more fundamentally user-oriented than the above is needed in developing index or search languages in new inter- and multi-disciplinary areas, where the documents may well originate in constituent disciplines, with their own conceptual or intellectual contexts and their own discrete languages. To provide useful and coherent access to such documents in their new context, new orientations and new language must be provided, by looking below explicit language to the association sets or concept systems with which the users in the new area organize their perception of the new area. Obviously to accept all the detail of all individual concept systems would lead to plethora and chaos, but even at the everyday level, Nunnally and Flangher (8) have shown that differences in word use, perception, personality, etc., resolve into some degree of consistency among people with similar background and interests.

In an unreported experiment in 1969, one of the present authors invited a group of documentation students to write down the five terms that each student associated most significantly with the term BOOK. The response from thirty students was well over seventy terms, and some sets of terms had no terms in common: at first a dismaying semantic diversity. However, many terms did recur, and some appeared frequently: LIBRARY, CULTURE, READING, CIVILIZATION. These words represented nodes in the generalized system of associations -- the terms accepted by the group as common concepts related to the trigger term.

Similar work in the College of Librarianship Wales in 1970 suggested extending the experiment by using the nodal terms as trigger terms for a second generation of responses, which could then be analyzed for further nodal terms in an expanded system. The intention in 1970 was to continue analysis and expansion in order to arrive at a generalized association set for the group in which the most frequently used and/or associated terms would appear. It was thought that such a generalized association set might even provide the basis for a user-oriented vocabulary. One of the present authors described a proposed system in 1973 (1) and reported on a pilot study in 1974 (2).

The process of eliciting and identifying a common and significant vocabulary from a user group can be conceptualized as a structured word association exercise, in which each member of the group responds to an initial trigger term by providing a list of terms that he/she closely associates with the trigger term. To arrive at a number of terms large enough to be valuable as an index language in the kind of context under discussion (perhaps 1000 terms), the user group must respond to several generations of trigger terms; in effect, they play several "rounds" of the exercise. Analysis of frequency of occurrence and pairing of terms in the user group's responses identifies the nodal terms used by the group. The most significant responses are used as the trigger terms for the next generation of responses; an algorithm for scoring the terms ranks their importance. A cut-off point for the exercise may depend on a number of factors: the size of the user group; the scope of the subject area; the size of a working vocabulary in that subject area (the last two factors are dependent but not identical); whether or not the vocabulary should be an immediate working vocabulary or a "seed" vocabulary for later development; and simply, the tolerance of the members of the user group. Ideally the exercise would continue until the appearance of new nodal terms is very infrequent; at this point it could be suggested that an operating vocabulary has been identified.

Even a user group as small as five people quickly produces a large volume of material. Research at the University of Maryland, College of Library and Information Services in 1980 was intended to determine the feasibility of using a computer to

perform much of the analysis necessary to carry out such an exercise. The initial software has been written to accommodate five users, each providing five responses per trigger term in each round. The one-fifth highest ranking terms from each round become the trigger terms for the next round. These parameters have been set in recognition of the limits of individuals to respond to geometrically increasing volumes of trigger terms. To illustrate, here are the number of terms and resulting trigger terms per round, for just five rounds:

1st round	1 trigger term requiring	5 responses
2nd	5	25
3rd	25	125
4th	125	625
5th	625	3125

Even so in practical terms the only humane way to generate responses is to ask for them in groups of five, spread out over time when necessary.

The significance of each term in the round is determined by assigning points for:

a) frequency of occurrence
b) pairing with other terms
c) relative distance from the trigger term (i.e., its ordinal position in the response to the trigger term)

In the following detailed description of the scoring system, each individual's five responses to a single trigger term are referred to as a "concept system":

STEP 1 Identify recurring terms. Score 1 point for each occurrence.

STEP 2a Identify position in concept system. Score:
5 points for 1st position
4 points for 2nd position
etc.

STEP 2b Identify recurring appearances with another term. (Two terms appearing together is called a "dyad" relationship.) Score:
4 points for being adjacent to the term
3 points for a 1 term separation
2 points for a 2 term separation
1 point for a 3 term separation

STEP 3a Weight individual terms in dyads which may have accumulated low scores because of low positions in STEP 2a. Weighting recognizes the relevance of dyads. Score:
5 points for 5th position
4 points for 4th position
etc.

STEP 3b Weight individual terms in triads which may also have accumulated low scores in STEP 2a. Triad relationships are recognized as having very high relevance. Score:
10 points for 5th position
8 points for 4th position
6 points for 3rd position
etc.

To illustrate the use of the scoring system, consider the following responses to the trigger term COAST:

Round 1 Trigger term: COAST

Response #1: COASTAL, ZONE, LITHOSPHERE, OCEAN, POLLUTION
#2: COASTAL, ZONE, EROSION, ESTUARY, MARSH
#3: ECOLOGY, COASTAL, ZONE, SHIFT, EROSION
#4: EROSION, COASTAL, LITTORAL, CURRENT, DEPOSIT
#5: ZONE, RESOURCE, SHIFT, DREDGE, EROSION

The step-by-step scoring is shown for each of the recurring words in this round:

	COASTAL	ZONE	EROSION	SHIFT
Step 1	4	4	4	2
Step 2a	18	16	10	5
Step 2b	23	27	24	17
Step 3a	12	19	32	18
Step 3b	12	22	36	24
Total Scores	69	88	106	66

(LITHOSPHERE, OCEAN, POLLUTION, ESTUARY, MARSH, ECOLOGY, LITTORAL, CURRENT, DEPOSIT, RESOURCE, and DREDGE occur once each and have a score of 1.)

Analysis shows that the scores for ZONE and EROSION are higher because they occur in several dyad and triad relationships. Relationships accumulate points at a higher rate than position in a concept system. COASTAL has high points for position, but low points for association. It finishes only slightly higher than SHIFT which had low position points, but appeared in two triad relationships.

Preliminary findings indicate that even such a basic scoring system requires a program capable of identifying several linguistic relationships among words, such as words having the same root, singular and plural forms of the same word, and common alterations in meaning caused by prefixes and suffixes.

The software written thus far has been written in SNOBOL 4 (9, 10, 11, 12), chosen because of its pattern matching capability. Other languages which have string variables functions may also be suitable for use, especially if they provide other valuable built-in functions.

The algorithm developed to analyze user responses has been designed to recognize relationships of terms within each individual concept system and also among all concept systems combined. SNOBOL's strengths in pattern matching and string manipulation have affected the program design significantly. For example, input data (all users' concept systems) are initially processed as one string. Each group of five terms can easily be manipulated as a string and searched for patterns and each term manipulated as a string as well.

The first operation performed on the data is the standardization of the spelling of response words so that similar forms of the same word are scored as the same word. This is done by grouping words with the same root together and identifying the most frequently occurring variation of the root word. Next, each term receives a position number tag designating its position within the concept system. Each term in the string is compared to the rest of the string and each time a match occurs, the frequency count for that term is increased. The frequency counts are stored in a table which associates a value with each term. All terms with frequency counts of 2 or more are stored in a string representing recurring terms.

Within each concept system, several new strings, representing all of the permutations of pairs of terms among the five terms, are created. Each member of each pair is compared to the recurring term string. If both members of a pair are found in the recurring term string, a significant relationship is indicated, and the score of each term in the pair is increased based on position in the concept system and distance from the other term. Scores are cumulated through analysis of the position number tags. A similar procedure identifies and scores triad relationships. Finally, the table containing scores for each term is converted to a two-dimensional array showing each term and its score. This array can then be sorted to identify the highest scoring terms, which can be used as the trigger term for the next round.

The scoring system has been applied so far to a relatively small sample, and work is continuing to refine it. Of particular interest is whether or not term relationships are truly more significant than frequency of occurrence.

Several other questions remain to be explored. The investigation so far has concerned the establishment of a common vocabulary by analysis of word frequency, relationship, and distance of association. However, relationships between terms may begin to emerge. For example, different terms consistently receiving the same group of associated terms are candidates for synonymous or quasi-synonymous relationship (USE and USED FOR). Further, since it is possible for a trigger term in an early round to appear as a response term to a different trigger term in a later round (especially when many rounds and terms are involved), some distinction may be made between coordinate and super-/sub-ordinate terms.

SUBTIL has several advantages in practical application. Its use of computer analysis of user vocabulary means that the real time involved in establishing a vocabulary is short; the elapsed time depends on the availability and resilience of members of the user group. The nature of the technique suggests its application to small areas of knowledge and small user groups, even though both knowledge and users may be homogeneous only in current concern and heterogeneous in origin. Thus, a highly relevant vocabulary could be developed quickly for a group that would not normally have one available and could not afford to develop one. Further, the technique would allow the vocabulary to develop and even change orientation as new members joined the user group. Such

a language could be used to index a new collection of documents appropriately for its users, to act as a basis for a concordance between user language and existing index languages, or even to offer assistance in text searching systems.

Finally, as one of the present authors has pointed out previously, the simpler and quicker the development of an index language, the less likely we are to become emotionally attached to it because of an investment of time and money (as seems to be the case with the conventional schemes of classification and subject headings). We need not then seek to maintain it beyond its useful life. It could become the recyclable thesaurus.

REFERENCES

1. BATTY, C.D. The instant thesaurus: synthesising users' concept systems to reduce cost and time in index language development. In Innovative developments in information systems: their benefits and costs; edited by Helen J. Waldron and F. Raymond Long. Proceedings of the American Society for Information Science, 1973. Vol. 10. pp. 11-12.

2. BATTY, C.D. Synthesized user based terminology index languages (SUBTIL). In Experience, problems, solutions: research and practice by Canadians; edited by André Gamache and Rudy Penner. Proceedings of the Second Open Conference on Information Science in Canada, 1974. pp. 109-114.

3. MOOERS, C.N. The indexing language of an information retrieval system. In Information retrieval today; edited by W. Simonton. Minneapolis: Center for Continuation Study, University of Minnesota, 1963. pp. 21-36.

4. BOURNE, C.P. Requirements, criteria, and measures of performance of information storage and retrieval systems, by C.P. Bourne, G.D. Peterson, B. Lefkowitz, and D. Ford. Palo Alto: Stanford Research Institute, 1961.

5. GREER, F.L. Word usage and implications for storage and retrieval. Washington, DC: General Electric Co., Information Systems Operations, 1962.

6. GREER, F.L. User vocabulary in thesaurus development. Perceptual and Motor Skills, 21 (1965), 827-837.

7. REISER, P. Evaluation of a growing thesaurus. Yorktown Heights: IBM Watson Research Center, 1966.

8. NUNNALLY, J.G. and FLANGHER, R.L. Psychological implications of word usage. Science, 140 (1963), 775-781.

9. GRISWOLD, R.E., PONGE, S.F., and POLONSKY, I.P. The SNOBOL4 programming language. 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1971.

10. GRISWOLD, R.E. and GRISWOLD, M.T. A SNOBOL4 primer. Englewood Cliffs, NJ: Prentice-Hall, 1973.

11. MAURER, W.D. The programmer's introduction to SNOBOL. New York: Elsevier, 1976.

12. NEWSTED, P.R. SNOBOL: An introduction to programming. Rochelle Park, NJ: Hayden Book Co., 1975.